North East Linguistics Society

Volume 29 *Proceedings of the North East Linguistic Society 29 -- Volume One: Papers from the Main Sessions*

Article 31

1999

Resolution of Scope Ambiguities in *How many* Questions

Elisabeth Villalta
University of Massachusetts, Amherst

Resolution of Scope Ambiguities in *How many* Questions

Elisabeth Villalta

University of Massachusetts, Amherst

1. Introduction

Theories of sentence comprehension study how perceivers understand sentences. One productive method has been to examine the strategies that comprehenders use in order to resolve ambiguity. This paper investigates the resolution of scope ambiguity in questions, that is, resolution of the ambiguity that arises when the wh-constituent interacts with other quantificational elements in the sentence. While scope ambiguity resolution in declarative sentences has recently received some attention in the psycholinguistic literature (e.g., Kurtzman & MacDonald (1993), Tunstall (1997)), this phenomenon, so far, has not been studied for questions. We will focus on the case of ambiguous *how many* questions that contain a universally quantified subject.

The two main contributions of this paper are the following. First, scope preferences in questions will be shown to be problematic for any economy-based approach to the processing of meaning, as indicated by the results of French and English questionnaire studies.

Second, a model will be developed in which these scope preferences are determined by the interaction with context. While incremental context interactive models have been claimed to induce immediate resolution of structural ambiguity (Crain & Steedman (1985), Altmann & Steedman (1988) and others), I will argue that the interaction with context can also be a reason to delay such ambiguity resolution. Evidence for this claim will be provided by a self-paced reading study in English.

The structure of this paper is as follows. In section 2, I present the phenomenon of scope interaction in *how many* questions. In section 3, I lay out the basic assumptions on the parsing model that I will use. I then go on, in section 4, to propose a first hypothesis for

Pius Tamanji, Masako Hirotani and Nancy Hall (eds.), NELS 29:443-457

scope ambiguity resolution in *how many* questions. This hypothesis is based on the assumption that the parser obeys some economy principle when constructing an LF representation of a sentence. The questionnaire studies in French and English, presented in section 5, provide strong evidence against the hypothesis. In section 6, a proposal is developed in which scope resolution derives from the interaction with context; and its hypothesis is tested in a self-paced reading study in English, in section 7.

2. Ambiguous *How many* Questions

In this section, I present the empirical facts about ambiguous *how many* questions. In a question, a particular interpretation is revealed by the answer that it requires in a given context. *How many* questions that contain a universally quantified subject allow for at least two interpretations. Let me illustrate this fact with a concrete example. Imagine the scenario described in (1).

(1) For the last semantics assignment, the students had to answer five questions out of seven and for each question a different paper had to be read. Thus, every student read (at least) five papers. Obviously, the students ended up reading different papers. Only two papers had been read by everybody.

In the context of the situation described above, the question in (2) allows for two possible answers, namely (2a) and (2b).

(2) How many papers did every student read for this last assignment?
   a. Answer 1: Five papers
   b. Answer 2: Two papers.

The question in (2) is a case of scope interaction between *n-many papers* and the subject quantifier *every student*. The two possible interpretations are given in (3a) and (3b).

(3) How many papers did every student read for this last assignment?

   a. For which n: $\forall$ x student(x)$\Rightarrow$ $\exists$Y papers(Y) & card(Y) = n & read(Y)(x)
   'For which number n: every student read n-many papers.' (Answer: Five papers)

   b. For which n: $\exists$Y papers(Y) & card(Y) = n & $\forall$ x student (x)$\Rightarrow$ read(Y)(x)
   'For which number n: there is a set of n-many papers that was read by every student.' (Answer: Two papers)

In what follows, I will assume a semantics for questions along the lines of Hamblin (1971) and Karttunen (1977), where the denotation of a question is the set of propositions which constitute (true) answers to that question. This approach requires a question operator Q in C^0 that turns a proposition into a set of propositions. I will follow Cresti (1995), by decomposing *how many N* into a *what n* part and a *n-many N* part. The latter can then be reconstructed into the position in which it gets interpreted while the former takes matrix scope as required by the question meaning. Two reconstruction sites for *n-many N* allow us

to postulate the two LF representations stated in (4a) and (4b).

(4) a. ***many*-downstairs** (every > many)

LF1: $[_{CP}$ How$_n$ $[C^0 [_{IP}$ every student$_i$ $[[$ t_n-**many papers**$]_j$ $[_{VP}$ t_i read t_j $]]]]]$

b. ***many*-upstairs** (many > every)

LF2: $[_{CP}$ How$_n$ $[C^0[_{IP}[$ t_n-**many papers**$]_j$ $[_{IP}$ every student$_i$ $[_{VP}$ t_i read $t_j]]]]]$

We will refer to the reading represented in (4a) as the *many*-downstairs interpretation because [t_n-*many papers*] is interpreted below the quantified subject, whereas the one in (4b) will be referred to as the *many*-upstairs interpretation. The downstairs reading in (4a) requires the 'Five papers'-answer, while the upstairs reading requires the 'Two papers'-answer, as can be seen from the corresponding paraphrases in (3a) and (3b).

After having illustrated the case of an ambiguous *how many* question, let me turn to a *how many* construction that does not display this ambiguity. French has a corresponding split construction, in which only *combien* ('*how many*') is fronted. This construction only allows for a *many*-downstairs interpretation. The question in (5) requires the 'Five papers'-answer in the scenario described above.

(5) **Combien** tous les étudiants ont-ils lu **de livres**?
How many all the students did-they read **of books**?

a. ✓ Answer 1: Five papers
b. * Answer 2: Two papers

The French non-split counterpart behaves like the English construction in that it is ambiguous and allows for both answers, as illustrated in (5').

(5') **Combien de livres** tous les étudiants ont-ils lu ?
How many of books all the students did-they read?

a. ✓ Answer 1: Five papers
b. ✓ Answer 2: Two papers

French thus presents the case where two different constructions can be used to express the same interpretation (the *many*-downstairs interpretation). While the split construction can only be used to express that particular reading, the non-split construction is ambiguous. I will argue, in section 4, that this configuration can be expected to have consequences for the strategies employed by the parser. Before doing so, in the following section, I turn to the underlying assumptions on the processing model adopted in this paper.

3. Assumptions on the Processing Model

Let me begin by laying out the basic assumptions on the parsing model that will be used here. In what follows, it will be assumed that, in order to compute an interpretation resulting from a particular scope configuration, the parser needs to build the corresponding LF representation. Therefore, when the scopal elements are not in the appropriate configuration at Surface Structure, the parser has to build the structural representation that can feed semantic interpretation. I will assume that the LF representation is computed along with the Surface Structure as the incoming words are perceived.

For the case of lexical ambiguity, it has been claimed that the parser can access all the different meanings of a particular lexical entry first and then choose the most appropriate meaning according to a number of principles. By contrast, for scope ambiguity, I will assume that ambiguity resolution does not require the comparison of several interpretations. Instead, I will propose that the different interpretations are calculated serially. The ambiguity being of structural nature, it will require the calculation of several LFs.[1]

The model of quantifier scope resolution presented here differs from the one proposed in Kurtzman & MacDonald (1993), where it is claimed that an analogy with lexical ambiguity resolution can in fact be made. In their model, alternative interpretations are initially considered in parallel. The single interpretation that best satisfies the scope principles is then selected.

Similarly, Crain & Steedman (1985) and Altmann & Steedman (1988) assume a model in which the different possible interpretations of a sentence are calculated in parallel. Their main claim is that structural ambiguity can be resolved early through incremental calculation of the interpretation and immediate consultation of the context. At the point of ambiguity, the reading is chosen that has less unsupported presuppositions with respect to the context (following their 'Principle of Parsimony'). Parallelism is claimed to be necessary, since such a decision can only be taken if several readings are compared. In their view, then, there is no absolute way to decide whether a sentence is the only possible or best continuation in a context.

As I will show below, in section 6, my proposal will share with this latter model the assumption that incremental calculation of meaning and early consultation of the context can play a role in the resolution of ambiguity, and in particular of scope ambiguity. However, I will propose that scope resolution in questions does not require comparison of all possible interpretations. Assuming a serial context interactive model as a possible alternative, I will argue that in certain cases the context can determine whether a reading is possible or not, without the need to compare different interpretations.

---

[1] Note that this approach has certain problems that cannot be addressed here. Notably, a serial account predicts that, if the construction of one LF is successful, there should be no reason to build more than one LF representation. Nevertheless, ambiguity can in fact be perceived. The alternative approach, in which all possible LFs are calculated in parallel, however, cannot explain why we do not always perceive ambiguity.

4. Attempting an Economy-based Approach

This section constitutes a first attempt to state a hypothesis about the order in which several LF representations are associated with an ambiguous *how many* question.

Following standard assumptions, the parser will be conceived as a mechanism that obeys principles of economy (cf. the Minimal Attachment Principle, Late Closure (Frazier,1978), Recency (Gibson, 1991), Simplicity (Gorrell,1995), the Minimal Chain Principle (De Vincenzi, 1991), and many others). Under such a perspective, it is reasonable to assume that the parser first chooses to construct the LF that requires minimal cost (e.g., the Principle of Scope Interpretation (Tunstall,1997) and the Minimal Lowering Principle (Frazier, 1997)). In particular, I will define the following cost function to make such a proposal explicit. Let us assume that the cost of an LF in which the quantifiers have been permuted is higher than that of the LF where the order of such elements is preserved (e.g., cost(Q1Q2)=0 < cost (Q2Q1)= 2). According to such a cost function, we expect the LF assigned first to respect the order of the quantificational elements in which they appear at Surface Structure. This idea is formulated in the Economy Hypothesis in (6).

(6) **Economy Hypothesis**
When processing a *how many* question, the parser first computes LF2, because it has less cost than LF1 (according to the cost function defined above).

Following this hypothesis, we expect the *many*-upstairs reading to be the preferred interpretation. For clarification, the two LFs are repeated in (7a) and (7b).

(7) a. *many*-downstairs (every > many)
LF1: $[_{CP}$ How$_n$ $[C^0$ $[_{IP}$ every student$_i$ $[$**[t$_n$ -many papers**$]_j$ $[_{VP}$ t$_i$ read t$_j$]]]]

b. *many*-upstairs (many > every)
LF2: $[_{CP}$ How$_n$ $[C^0[_{IP}$ **[t$_n$-many papers**$]_j$ $[_{IP}$ every student$_i$ $[_{VP}$ t$_i$ read t$_j$]]]]]

In what follows, I want to argue that, in the French case, the preference for the upstairs reading is expected to be even stronger. Because French allows for two constructions that differ with respect to their possible scope configurations, a particular version of the Blocking Principle (cf. Aronoff (1976)) can be argued to apply. Before proposing a second hypothesis, let me briefly introduce this principle and show how it applies to the French data.

The Blocking Principle has been claimed to hold in the lexicon (i.e., certain words do not exist because other words with identical semantics do). Williams (1997) proposes that this principle can also be extended to other levels of the grammar, in particular to syntax. Williams (1997) interprets Aronoff's (1976) principle in the following way: "if two forms exist (in syntax or morphology), they must have different meanings" (p.578). Similarly, in psycholinguistics, there has been an attempt to use the Blocking Principle to account for certain attachment preferences. Frazier & Clifton (1997) explain the different relative clause attachment preferences in English and Spanish with a Gricean version of this principle.

It is possible to extend the Gricean version of the Blocking Principle to the French data in the following way. The existence of the unambiguous split construction should 'block' the ambiguous non-split construction from receiving the downstairs interpretation. In fact, this configuration disconfirms Williams' initial proposal. His proposal predicts that the non-split construction should not allow for ambiguity at all. The unambiguous split construction should prevent the non-split construction from receiving the downstairs interpretation. A Gricean version of this principle can however apply: we expect the non-split construction to be used more often to express the upstairs reading than the downstairs reading. Consequently, the parser should prefer to assign the upstairs reading to the non split construction. This hypothesis is stated in (8).

(8) **Blocking Hypothesis**
In French, the split question is unambiguously used to express the *many*-downstairs reading. Assuming a Gricean version of the Blocking Principle, we expect the non split question to be used in production to express the *many*-upstairs reading, and correspondingly expect perceivers to favor that reading in comprehension.

Both hypotheses, (6) and (8), predict a preference for the *many*-upstairs interpretation. To test these two hypotheses, I carried out two questionnaire studies in English and in French. The preferred interpretation of a *how many* question was determined with the help of the answers that participants chose in particular contexts. These questionnaire studies are presented in the following section.[2]

5. Experiment I: French and English Questionnaire Studies

5.1. Method

Participants read twelve stories followed by a question. The eight experimental stories were designed to equally support both interpretations and were followed by a *how many* question. Participants were asked to write down the first answer that came to mind. They were also encouraged to indicate other correct answers when possible. In the French questionnaires, the split/non-split nature of the question was manipulated. In the English questionnaires, the partitive/ non-partitive nature of the *how many*-phrase was manipulated.

(9) Example:

Three friends went to the last Music Festival in Montreal. Altogether, each of them saw ten movies. When comparing what they had seen at the end, they realized that there were four movies that they all had seen.

How many (of the) movies did everybody see at the Movie Festival in Montreal?

[2] For reasons of space, the experiments presented in this paper cannot be described in all details. The rather succinct description of the experiments, however, contains all the information relevant to the argumentation in this paper.

5.2. Participants

32 UMass undergraduate students and 32 undergraduate students from the Université Paris 8 participated in this experiment. The UMass undergraduate students received extra course credit for it; the French students completed the experiment as part of a classroom exercise.

5.3. Results

The results disconfirm the prediction of the hypotheses (6) and (8), as can be seen in Tables 1 and 2.[3] Both in English and French, the questions were answered more often with a downstairs interpretation than with an upstairs interpretation. No significant difference was found between partitive and non-partitive questions nor between split and non-split questions. Some of the questions were answered with a cumulative answer (the total number of sets mentioned in the previous discourse).

T-test results: in French, there was a highly significant preference by subject ($t_1(31) = 9.67$, $p<.0001$) and by item ($t_2(7) = 11.88$), $p<.0001$). In English, there was a significant preference by subject ($t_1(31)=2.75$, $p<.007$), but not by item ($t_2(11) = 1.59$, $p<.15$). However, seven out of the eight English stories had a significant preference for the downstairs interpretation (sign test: $p<.035$).

Table 1
Number of answers (French)

	Upstairs	Downstairs	Cumulative	Total
Split	5	105	12	126
Non split	9 (+4)	107 (+1)	12	128
Total	**14 (5.5%)**	**212 (83.5%)**	**24**	**254**

Table 2
Number of answers (English)

	Upstairs	Downstairs	Cumulative	Total
Non partitive	46 (+1)	74 (+4)	7(+1)	127
Partitive	50 (+1)	74	3	127
Total	**94 (37.8%)**	**148(58.3%)**	**10**	**254**

[3] In Tables 1 and 2, the numbers in parentheses indicate second choice-answers.

5.4. Discussion

The two questionnaire studies provide evidence against the Economy Hypothesis. The results suggest, in English and in French, that the first LF computed is not LF2, the LF that has less cost, but rather LF1.[4] Furthermore, the French questionnaire study provides evidence against the Blocking Hypothesis. The preference for a downstairs interpretation in a French non-split construction is as strong as in a split construction, contrary to expectation.

Before turning to an alternative hypothesis, in section 6, let me point out that this result is unexpected under the current theories on the processing of meaning, in particular, under the Immediate Interpretation Principle widely adopted in on-line interactive approaches (cf. Marslen-Wilson & Tyler (1980), Crain and Steedman (1985), Altman & Steedman (1988) and others). In these approaches, the interpretation of an utterance is computed immediately and in an incremental fashion. Extensive experimental evidence has shown that the processor commits to a particular word meaning at the earliest point possible and that this has immediate consequences for the representation of the input. Furthermore, in such models, the processing of an utterance is claimed to always be conducted with immediate reference to the discourse context in which it occurs. Crain and Steedman (1985), for example, argue that "the primary responsibility for the resolution of local syntactic ambiguities in natural language rests not with structural mechanisms, but rather with immediate, almost word-by-word interaction with semantics and reference to the context." (p.321)

Note, however, that the Immediate Interpretation Principle predicts a preference for the upstairs interpretation of a *how many* question. Immediate Interpretation has as a consequence that each incoming word is interpreted immediately and is therefore integrated into the LF representation as soon as possible. We thus predict that the parser should first attempt to construct LF2. The constituent *n-many N* should be interpreted and integrated into the LF as soon as it is encountered. Hence, the parser is committed to the upstairs interpretation as soon as the *how many* phrase has been processed. For clarification, the two LFs are repeated again in (10a) and (10b).

(10) a. *many*-downstairs (every > many)
LF1: $[_{CP}$ How$_n$ $[C^0$ $[_{IP}$ every student$_i$ $[[$ t_n **-many papers**$]_j$ $[_{VP}$ t_i read t_j $]]]]$

b. *many*-upstairs (many > every)
LF2: $[_{CP}$ How$_n$ $[C^0[_{IP}$ **[** t_n**-many papers**$]_j$ $[_{IP}$ every student$_i$ $[_{VP}$ t_i read $t_j]]]]]$

In the next section, I present a model in which interaction with context can also be a reason to delay the interpretation and LF-integration of a constituent. As a consequence, the more 'economic' LF1 does not necessarily have to be the first LF constructed by the parser.

---

[4] These results are challenging for the current scope resolution and LF-principles used in the psycholinguistic literature (cf. Kurtzman & Mac Donald 1993, Tunstall 1997), which cannot account for this preference. Although these theories do not claim to have an account for questions, a unified analysis of quantifier scope resolution in declaratives and questions would be desirable.

6. Scope Ambiguity Resolution determined by the Interaction with Context

In what follows, I will adopt a model in which the parser can access information from the discourse context. Similarly to what is proposed in Crain & Steedman (1985) and Altmann & Steedman (1988), I will assume that resolution of local structural ambiguities depends on the interaction with semantics and reference to context.

A discourse model can be represented as a conversational record (cf. Stalnaker 1969). As already described in Karttunen (1970), it can be viewed as a file that consists of records of all the individuals mentioned in the text, and for each individual of a record that contains its properties. In particular, I will assume, following Heim (1982), that indefinites introduce new discourse referents into the representation, while definites and pronouns have to refer back to antecedents introduced earlier into the discourse. Furthermore, I will claim that wh-phrases also require an antecedent in the discourse. In the particular case of a *how many* phrase, the search for its antecedent-set in the discourse is done in order to determine its cardinality.

In the previous section, I concluded that the Immediate Interpretation Principle commits the parser to the upstairs interpretation, as soon as the *how many* constituent has been encountered. This is a consequence of the fact that incoming elements are, under the Immediate Interpretation Principle, interpreted and integrated into the LF as soon as possible. Nevertheless, I want to argue here that within a context interactive model, immediate access to context can as well be a reason to delay the decision to integrate a particular element into the LF representation. I want to claim that, under a slightly different formulation, the Immediate Interpretation Principle is in fact not incompatible with the preference to interpret a *how many* question with a downstairs interpretation.

Let me now illustrate, for the particular case of a *how many* phrase, how information from the context can delay its integration into the LF-representation. The lexical information from this constituent triggers the parser to search for a set in the discourse (an antecedent for *n-many N*). Immediate Interpretation predicts that this search should be done as soon as possible, that is, as soon as the constituent has been processed. The wh-phrase by itself, however, does not necessarily have enough descriptive content to allow the parser to choose a particular antecedent from the discourse. If there is more than one possible antecedent in the discourse, the parser cannot choose the appropriate antecedent for *n-many N*. Only if the context contains a unique salient antecedent can the search be successful and an interpretation be assigned immediately to the constituent. We can conclude that only particular contexts allow wh-phrases to be interpreted immediately. I will take 'assigning an interpretation to a constituent' to mean 'successfully integrating it into the LF representation'.

If there are several possible antecedents in the discourse, the parser cannot integrate *n-many N* into the LF representation immediately. I will assume that the constituent is, at that point, put in a stack. It remains in the stack until there is enough information available to choose its appropriate antecedent. Only then can it be integrated into the LF representation. A hypothesis for the scope resolution in *how many* questions is formulated in (11).

(11) Context Dependency Hypothesis
How many N triggers the search for an antecedent-set in the discourse. A context that provides more than one possible antecedent for *n-many N* delays its incorporation into the LF representation.

There is a parallelism with the discourse conditions that pronouns/definite descriptions require for their antecedents.[5] For illustration, consider the following examples, taken from Heim (1982).

(12) John has a cat and a dog. ?Its name is Felix
(13) John has a cat and a dog. ?The pet's name is Felix

Heim (1982) shows that the use of an anaphoric element is only acceptable if it is clear which of the discourse referents in the context it refers to. In (12) and (13), there are two equally possible antecedents for the pronoun *it* and the definite description *the pet*. The lack of enough descriptive content makes it impossible to find the appropriate antecedent. Similarly, Kadmon (1987) proposes a general uniqueness condition on definite descriptions and pronouns. These are only felicitous if they have a unique salient antecedent in the context. I assume here that *wh*-phrases are anaphoric, and therefore they also obey the uniqueness condition. They can only be interpreted felicitously if there is a unique salient antecedent in the discourse context.

The Context Dependency Hypothesis makes the following predictions. If a context contains a single antecedent for *n-many N*, the parser should commit to the upstairs interpretation immediately. Under this condition, we expect a preference to answer the question with an upstairs interpretation. This prediction is not incompatible with the results of Experiment I. In the questionnaire studies, the stories preceding the questions were designed to equally support both interpretations. None of the sets mentioned in the discourse was particularly salient.

If a context contains more than one salient antecedent for *n-many N*, we predict that this constituent should not be immediately incorporated into the LF representation. Therefore, the parser does not commit to the upstairs interpretation immediately. We predict that questions should receive less upstairs answers than in contexts that contain a unique salient antecedent. We expect, furthermore, this delay in interpretation to be reflected in the reading times of the question. There should be a difference in processing load at the point at which *n-many N* is integrated into the LF representation. Presumably, once the verb has been encountered, enough information is available to take this decision. At that point, we expect higher reading times than in contexts that contain a unique salient antecedent.

In order to test the predictions of the Context Dependency Hypothesis, I conducted a self-paced reading study in English, in which participants had to read questions on a computer screen preceded by a context. The self-paced method allowed us to measure the

[5] These conditions are discussed in Heim (1982) and Kadmon (1987).

processing load of the different regions of the question and to determine whether a difference in context affected the reading times. The experiment is described in the following section.

7. Experiment II: Self-paced reading study in English

7.1. Method

Participants read sixty-six stories on a computer screen followed by a question or continuation sentence (self-paced method). Participants used a response key to indicate when they had finished reading the portion of text presented on the screen. Once the whole story was read, it disappeared from the screen. Each of the twelve experimental stories was followed by a *how many* question, which was presented region by region. The regions disappeared when the participant pulled the response key to indicate that s(he) had finished reading, and the following region appeared. The question was divided into five presentation regions: How many N / Quantifier / Verb (+ Particle) / Modifier 1/ Modifier 2. Once participants finished reading, the question disappeared from the screen and they were asked to choose one of the two answers presented on the screen (upstairs and downstairs answer), by pulling the corresponding trigger.

The presence/absence of a unique salient antecedent in the context was manipulated. Condition 1 presented the same stories from Experiment I, which supported both interpretations equally. Condition 2 presented the stories from Experiment I, but minimally changed. These supported the upstairs interpretation. One or two sentences were added to the original stories, in order to increase the salience of the set corresponding to the upstairs interpretation. This was done by involving the set in an additional event. The salience of the downstairs information was reduced by using a vague cardinality (*different, several, some*). Six stories presented the upstairs information before the downstairs information, the other six stories presented the downstairs information before the upstairs information.An example is given below.

(14) Condition 1
In December, the chef distributed some of his recipes to his students. There was one recipe that everybody received, the "Chilled Terrine with Pistachios and Caper Mustard". Altogether, each of them received four different recipes.

(15) Condition 2
In December, the chef distributed some of his recipes to his students. There was one recipe that everybody received, the "Chilled Terrine with Pistachios and Caper Mustard". That was his special recipe. He wanted to make sure that everybody would be able to try it out.

(16) Question following the context
How many recipes/ did every student/ receive from/ the chef/ in December?

7.2. Participants

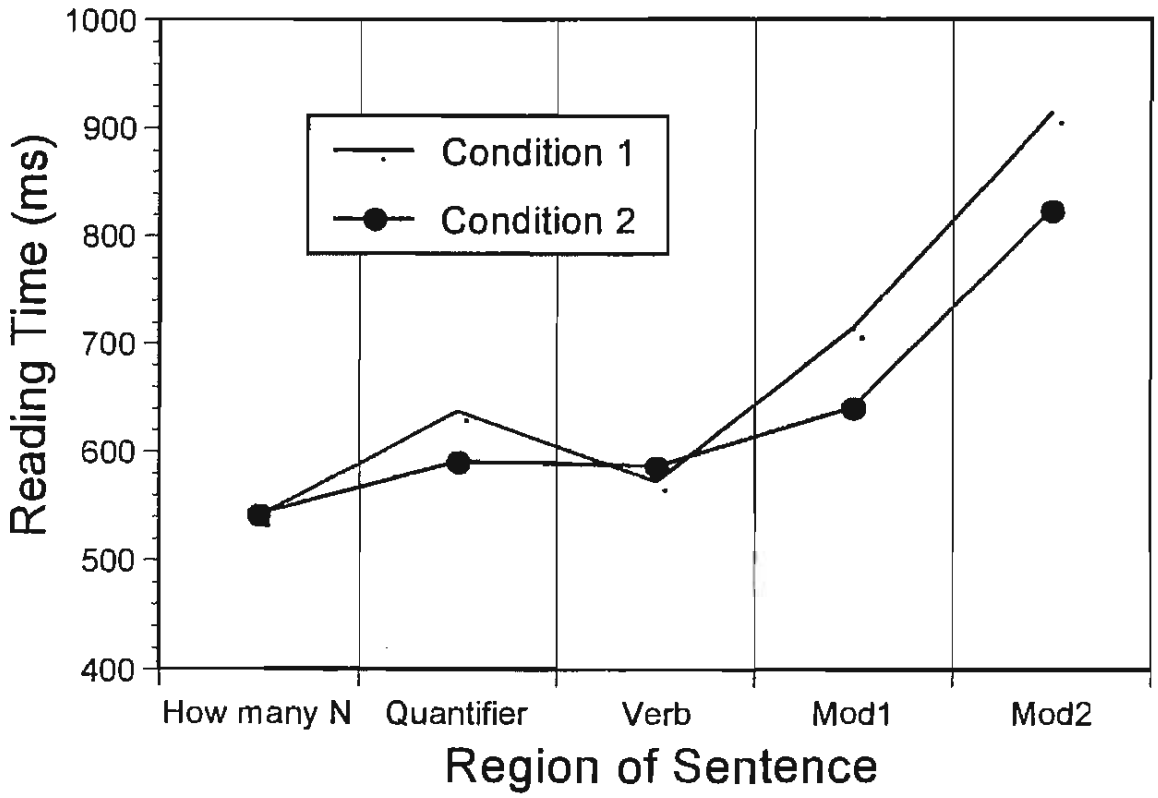
44 UMass undergraduate students participated in this experiment and received extra course credit for it.

7.3. Results

As predicted by the Context Dependency Hypothesis, in condition 1, participants were significantly slower in the last two regions (following the verb); the results were significant by subject, $F_1(43)=5.01$, $p<.03$, and nearly significant by item, $F(11) = 4.58$, $p<.06$. Figure 1 presents the reading times for both conditions associated with each of the five regions.

Condition 1 received significantly more downstairs answers than condition 2. There were 66% downstairs answers in condition 1 and only 47% downstairs answers in condition 2 ($F_1(43) = 10.13$, $p<.003$, $F_2(11)= 7.73$, $p.<.02$).

Figure 1



7.4. Discussion

The results of this experiment provide evidence that context affects the processing of a *how many* question. They indicate that immediate access to context does not require the parser to resolve ambiguity as soon as possible. The delay in the resolution of scope ambiguity is reflected in the fact that participants are significantly slower in reading the last two regions when the context does not provide a unique salient antecedent. Hence, immediate access to context can also be a reason to delay this ambiguity resolution. Immediate access to context does not have as a consequence that the parser necessarily commits to the upstairs interpretation immediately.

The fact that participants were also slower in the second region in condition 1 (but not significantly slower: $F_1(43)=3.06$, $p<.09$, $F_2(11)=1.42$, $p.3$) could be argued to support an alternative hypothesis, namely that the contexts in condition 2 induced a general lower processing load for the whole question. Since these only support the upstairs information, one could claim that the following question is generally 'easier' to process. However, if some general difficulty of the context in condition 1 slowed reading in both region 2 and region 4 for certain items, then one would expect reading time to correlate across items between region 2 and 4. No correlation was found (Pearson $r = 0.079$). We can conclude that contexts with a unique salient antecedent did not necessarily have as a consequence a general lower processing load, but that only the last two regions of the question (following the verb) were affected in a significant way.

This experiment has allowed us to determine that context can delay the resolution of quantifier scope ambiguity. Nevertheless, an open question still remains, namely, why delaying the interpretation of *n-many N* has as a consequence that the downstairs interpretation is preferred. This preference does not follow directly from the delay. It can however be argued that, under economy, the parser conforms to a structure preservation principle. Hence, in the case in which ambiguity resolution is delayed, there should be a preference to not give up the structure that has been successfully built up to that point. It follows that there should be a preference to construct the LF corresponding to the downstairs interpretation under these conditions.

8. Conclusion

The experimental results presented in this paper confirm that there is a preference to interpret a *how many* question with a *many*-downstairs interpretation when the preceding context supports the upstairs and the downstairs interpretations equally.

These results disconfirm the Economy Hypothesis formulated in section 4, since the preferred interpretation does not correspond to the more economic LF representation. These results are also unexpected under the current assumptions on the processing of meaning. Immediate Interpretation, a principle which has been widely adopted in on-line interactive approaches, does not seem to make the correct predictions. Under the current view of Immediate Interpretation, the parser should commit to the upstairs interpretation very early in the sentence. Therefore, the downstairs reading should be a very difficult reading.

The main point of this paper has been to show that, once context is taken into consideration, these results are in fact not incompatible with Immediate Interpretation. I have proposed a model in which scope ambiguity resolution is determined by the interaction with context. An important feature of this model is that the interaction with context can also be a reason to delay ambiguity resolution. Crucially, then, immediate access to context does not necessarily commit the parser to the upstairs interpretation. Only contexts that provide a single salient antecedent have this consequence.

The preference for the downstairs interpretation has been shown to be even more surprising in the case of French. A Gricean version of the Blocking Hypothesis predicts, for the non-split construction, a preference for the upstairs interpretation. However, the experimental results indicate that both split and non-split constructions received a preference for the downstairs interpretation. The effects of the Blocking Principle must be overridden by the requirements that context imposes on the interpretation of a *how many* question.

That context plays an important role in the resolution of ambiguity is a well known fact. The experimental results presented here have shown that this is also clearly the case for quantifier scope ambiguity in questions. In particular, I have claimed, that different contexts can directly affect how a question is processed. I conclude that a better understanding of the processing of quantifiers can only be reached if serious consideration is given to the contexts in which these appear.

References

Altmann, G. and M. Steedman. 1988. Interaction with context during human sentence processing. *Cognition* 30: 191-238.

Crain, S. and M. Steedman. 1985. On not being led up the garden path: the use of context by the psychological syntax processor. In *Natural Language Parsing*, ed. D. Dowty et al., 320-358. Cambridge: Cambridge University Press.

Cresti, D. 1995. Extraction and Reconstruction. *Natural Language Semantics* 3 : 79-122.

Frazier, L. 1997. On Interpretation: Minimal lowering. Ms., University of Massachusetts at Amherst.

Frazier, L. and C. Clifton. 1997. *Construal*. Cambridge: MIT Press.

Hamblin, C.L. 1971. .Questions in Montague English. *Foundations of Language* 10: 41-53.

Heim, I. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. PhD dissertation, University of Massachusetts at Amherst.

Kadmon, N. 1987. *On Unique and Non-Unique Reference and Asymmetric Quantification.* PhD dissertation, University of Massachusetts at Amherst.

Karttunen, L. 1976. Discourse Referents. In *Syntax and Semantics 7*, ed. J. McCawley, 363-385. New York: Academic Press.

Karttunen, L. 1977. Syntax and Semantics of Questions. *Linguistics and Philosophy* 1: 3-44.

Kurtzman, H. and M. Mac Donald. 1993. Resolution of Quantifier Scope Ambiguities. *Cognition* 48: 243-279.

Marlson-Wilson, W. and L. Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 8: 1-71.

Stalnaker, R. 1979. Assertion. In *Syntax and Semantics 9 - Pragmatics*, ed. Cole P, 315-332. New York: Academic Press.

Tunstall, S. 1997. *Interpreting Quantifiers.* PhD dissertation, University of Massachusetts at Amherst.

Williams, E. 1997. Blocking and Anaphora. *Linguistic Inquiry* 28,4: 577-627.

Department of Linguistics
South College
University of Massachusetts
Amherst, MA 01003

villalta@linguist.umass.edu